DOCUMENT RESUME

ED 355 275 TM 019 629

AUTHOR Aycock, Tim
TITLE It Is Incorrect To Say "The Test Is Reliable": A Review of the Literature and Implications for Research Practice.
PUB DATE Jan 93
NOTE 15p.; Paper presented at the Annual Meeting of the Southwest Educational Research Association (Austin, TX, January 28-30, 1993).
PUB TYPE Information Analyses (070) -- Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Educational Practices; *Educational Research; *Estimation (Mathematics); *Interrater Reliability; Literature Reviews; Meta Analysis; *Research Methodology; Research Reports; *Scholarly Journals; Test Construction; *Test Reliability; Trend Analysis
IDENTIFIERS *Journal of Counseling Psychology

ABSTRACT

To determine trends in reporting test reliability, 88 articles addressing 188 instruments in 1980, 81 articles covering 205 instruments in 1985, and 67 articles assessing 195 instruments in 1990 in the "Journal of Counseling Psychology" were reviewed. Articles were examined for the way in which reliability was discussed and reported, and were grouped into the following categories: (1) calculation of reliability estimates for the sample; (2) calculations of the reliability estimates for the sample when the instrument was developed or modified by the researcher; (3) calculations of inter-rater reliability; (4) general acknowledgment of an instrument's reliability without any specific estimated mentioned; and (5) failure to report any kind of reliability estimates. As expected, during the last decade, the general discussion of reliability appeared to increase. In 1980 reliability reports were not given for 55 percent (104) of the instruments used, while in 1990, only 24 percent (47) did not report this information. While authors' tendencies to discuss reliability estimates are growing, the reliance on past estimates appears to be strong. Analysis suggests that fluid views of reliability appear to be increasing, but static views seem to be the prevalent perspective. An upward trend is seen in the calculation of sample reliability. It is hoped that the trend in reporting estimates of reliability for each sample will increase the clarity of findings by limiting confounding factors. (SLD)

**********
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
**********

ED355275

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Tim Aycock

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

# It is Incorrect to Say "the Test is Reliable": A Review of the Literature and Implications for Research Practice

Tim Aycock

Texas A&M University

Paper presented at the annual meeting of the Southwest Educational Research Association, Austin, TX January 29. 1993.

TM019029

BEST COPY AVAILABLE

# It is Incorrect to Say "the Test is Reliable": A Review of the Literature and Implications for Research Practice

Psychometric instruments are used frequently in counseling psychology (Dawis, 1987; Meier & Davis, 1990), but despite their prevalent use, the full disclosure of the psychometric properties of given scales is often lacking. In their review of the Journal of Counseling Psychology Meier and Davis (1990) state that the majority of scales used were not accompanied by reports of the psychometric properties of the scales. Willson's (1980) review of articles in the American Educational Research Journal from 1969 to 1978 found that reliability estimates were only reported in less than half of the published articles. Willson (1980) described this unreporting as "inexcusable" and encouraged editors and reviewers to "routinely return papers that fail to establish psychometric properties of the instruments they use" (p. 9).

In an attempt to explain the unreporting phenomenon, Meier and Davis (1990) suggest that researchers may omit reporting psychometric data because no accepted standards for reporting scale information exist. They also conjecture that a belief persists among researchers that "psychometric properties of scales is technically important, [but] practically it is often of little use" (Meier & Davis, 1990, p. 115). Distorted responses, lack of theory, and low predictive validity are described as problems of psychology that have contributed to the difficulty in obtaining accurate measurements over the last 60 years.

In addition to the anesthetic effect of its common occurrence, disregarding information about instruments' dependability and validity can be dangerous to the extent that entire research endeavors may be invalidated. For example, Kazdin (1980) suggests that ignoring an instrument's sensitivity may become troublesome when attempting to interpret the results of an investigation. He states:

> If no relationship were demonstrated between the independent and dependent variables at the end of the investigation, it would be reassuring to know that the reason for this was not the insensitivity of the dependent measure. A common tactic among researchers is to lament the lack of sensitivity of a dependent measure as a possible explanation for the lack of findings. (Kazdin, 1980, p. 221)

Reliability coefficients also influence the ability to detect effect sizes. As Locke, Spirduso and Silverman (1987) state, "the correlation between scores from two tests cannot exceed the square root of the product for reliability in each test" (p. 28). The implications of these limiting effects are exposed by Snyder, Lawson, Thompson, Stricklin, and Sexton (in press, emphasis added):

> Reliability coefficients for the data obtained on study instruments used in the empirical investigation prospectively provide a basis for determining, a priori, whether a proposed study and substantive analyses are even plausible. These coefficients also allow the researcher to retrospectively interpret obtained effect sizes (e.g., $r^2$) against the ceiling created by the reliability coefficients obtained in a study.

While both validity and reliability are important components of measurement theory (Nunnally, 1982), issues of reliability are often misunderstood. Technically, reliability or dependability of measurements results when individuals' observed scores from an instrument are highly correlated with the individuals' "true" scores (Allen & Yen, 1979). Because the existence of "true" scores are assumed and usually cannot be proven, reliability can only be estimated. According to Sax (1980), reliability estimates can be increased when measurements provide consistent, unambiguous information. He states:

> Measurements are reliable if they reflect "true" rather than chance aspects of the trait or ability measured. To the extent that chance or random conditions have been reduced, reliability will be high, and measurements will provide dependable knowledge. Chance factors included conditions within the examinee (fatigue, boredom, lack of motivation, carelessness), characteristics of the test (ambiguous items, trick questions, poorly worded directions), and conditions of scoring (carelessness, disregard of or lack of clear standards for scoring, and counting and computational errors." (Sax, 1980, p. 255-256)

Melancon and Thompson (1990) note that reliability is often mistaken to be an inherent trait as evidenced by a test being described as reliable. In fact it is incorrect to say that "the test is reliable." Tests are not reliable or unreliable; rather, data have these characteristics, "albeit data generated on a given measure

administered with a given protocol to given subjects on given occasions" (Eason, 1991, p. 84). As Thompson (1992) notes:

> This is not just an issue of sloppy speaking--the problem is that sometimes we unconsciously come to think what we say or what we hear, so that sloppy speaking does sometimes lead to a more pernicious outcome, sloppy thinking and sloppy practice. (p. 436)

Sax (1980) also emphasizes the need to be precise when describing reliability by saying:

> ...it is more accurate to talk about the reliability of *measurements* (data, scores, and observations) than the reliability of tests (questions, items, and other tasks). Tests cannot be stable or unstable, but observations can. Any reference to the "reliability of a test" should always be interpreted to mean the "reliability of measurements or observations derived from a test." (Sax, 1980, p. 261)

Dawis (1987) agrees that "reliability is a function of sample as well as of instrument, [and] it should be evaluated on a sample from the intended target population--an obvious but sometimes overlooked point" (p. 486). Rowley's (1976) discussion of reliability in observational research also echoes this refrain:

> writers in the area have not made sufficiently clear to their readers that reliability is a property of a measure...and not of an instrument, or of a record. It needs to be established that an instrument itself is neither reliable or unreliable--it is only when the instrument has been used to collect data, and when the data have been manipulated in some way to produce

> scores, that we can speak sensibly about reliability. A single instrument can produce scores which are reliable, and other scores which are unreliable. (p. 53)

Researchers' operational definitions of reliability influence their use of psychometric measures. Those who perceive that reliability is a trait of the instruments employed may be more likely to rely on reliability estimates of the instruments from previous data sets. The following excerpts exemplify reliability reporting from this perspective:

> This instrument [the Counselor Research Form] has been found to be a reliable and valid means for measuring these perceived counselor characteristics. Split-half reliabilities of .87 for expertness, .85 for attractiveness, and .91 for trustworthiness have been reported. (Robbins & Haase, 1985, p. 508);
>
> Novaco (1977) reported a split-half reliability coefficient (Cronbach alpha of .96). (Hains & Szyjakowski, 1990, p.80);
>
> Research (Gottfredson, Holland, & Holland, 1978) on the scales measuring the Holland types for the seventh revision indicated acceptable reliability (Kuder-Richardson 20 values ranging from .85 to .91 for men and women.... (Slaney, 1980, p. 123)

The reliance of these authors on previous estimations of reliability requires some implicit assumptions. One assumption is that the observations made in their own studies were highly similar to the observations obtained when the reliability estimates were obtained. The difficulty of this assumption is that the original conditions

present in an instrument's development may be difficult to discern and even more difficult to replicate.

When a disparity is known to exist between a research sample and the sample on which an instrument was developed, the need for obtaining reliability estimates for the data in hand seems compelling. Long (1990) demonstrates this logic when she recognized that the sample in her study had a different demographic composition than the sample used to develop the Ways of Coping Checklist. She wisely computed the internal consistency for her sample. Perhaps determining the reliability estimates of any given sample should be considered routine when using instruments. Calculating reliability estimates may be a parsimonious and objective alternative to comparing a given sample to the sample the author used in evaluating the instrument.

Even if a researcher's sample is identical to the sample used to norm an instrument, a static view of reliability would require the researcher to assume that the participation of subjects were equivalent. This assumption seems dangerous because of the varying conditions under which subjects participate in and researchers conduct studies. For example, imagine the variability that could exist in the self-report of freshmen students in introductory courses. Investigators would seem to be accepting a big risk when the response sets of subjects (whose motivation may be suspect) are used without checking the dependability of those scores. Some instruments attempt to correct for this factor by providing validity scales and checks for random response sets, while others assume

subjects' participation is as consistent and valid as the sample used to norm the instrument.

Obtaining estimates of reliability on samples would appear to be an important step in increasing the clarity of research findings. Given the increased attention to issues of reliability, one might expect that the literature would reflect this increased interest. Even though the limitations of classical test theory are noted (Shavelson & Webb, 1991), it would still seem desirable for researchers to provide some kind of sample estimates of reliability as against providing only the citation of past sample estimates. Not mentioning specific reliability estimates of any kind would seem to be the least desirable response.

In response to these expectations the following hypotheses were made about the discussion of reliability in the last decade:

1. Reliability reporting will have risen.
2. There will have been more reports of reliability which relied on past estimates of reliability as opposed to current reliability estimates of the researchers' sample data.
3. There will have been an increase in the calculation of reliability estimates when instruments that have been previously established are used.
4. The reporting of reliability estimates with new or modified scales and inter-rater reliabilities will have been greater than the reporting of reliability estimates of sample data when established instruments were used.

## Procedure

To determine the trends of reporting reliability, the 1980, 1985, and 1990 articles in the Journal of Counseling Psychology were reviewed. The articles were examined for the manner in which reliability was discussed and reported. Distinctions were made among reports of reliability estimates along the following dimensions: calculation of reliability estimates for the sample; calculations of the reliability estimates for the sample when an instrument was developed or modified by the researcher; calculations of inter-rater reliability; the general acknowledgment of an instrument's reliability without any specific estimates mentioned; and the failure to report any kind of reliability estimates.

These categories were devised to aide in determining the authors' perspective on reliability estimates. Several assumption were made. Authors who provided sample reliability estimates even when established instruments were used were assumed to view reliability as sample dependent as opposed to instrument dependent. When reliability estimates were obtained in studies that employed new or modified instruments, a fluid perception of reliability was not necessarily assumed. It is possible for researchers who modify scales to obtain reliability estimates because they believe established instruments were tampered with and the original reliability estimates could potentially be contaminated.

Inter-rater reliabilities were identified separately because the direct involvement of human raters would seem to clearly identify a need to evaluate the reliability of those judges. While written self-report instruments may be assumed to produce reliable

measurements, instruments that require human raters would seem to expose the transient nature of reliability; however, it seems possible that researchers could identify inter-rater reliability as being sample dependent but still may not perceive other written instruments as susceptible to this variability.

## Results

Table 1

| | 1980 | 1985 | 1990 |
|---|---|---|---|
| Total no. of articles | 88 | 81 | 67 |
| Total no. scales | 188 | 205 | 195 |
| Citation of past reliability estimates | 35 *19%* | 41 *20%* | 82 *42%* |
| Reliability estimates on samples for scales developed or modified by investigator | 5 *3%* | 21 *10%* | 21 *11%* |
| Reliability estimates for samples using inter-rater reliability reported | 25 *13%* | 28 *14%* | 6 *3%* |
| Reliability estimates on samples for previously developed scales | 9 *5%* | 27 *13%* | 36 *18%* |
| General report of past reliability estimates | 10 *5%* | 22 *11%* | 3 *2%* |
| No reliability reported | 104 *55%* | 66 *32%* | 47 *24%* |

## Discussion

As expected, during the last decade the general discussion of reliability appeared to increase. Researchers appear to be reporting basic information about reliability. Perhaps warnings such as those given by Willson (1980) are gradually being heeded. In 1980 reliability reports were not given for 55% (104) of the instruments used, while in a 1990 only 24% (47) of the instruments were used without reporting this information. Part of this increase may be a reflection of the specific reporting of past reliability estimates, which

in 1980 occurred 19% of the time and increased in 1990 to 42% of the time. While there are limitations to viewing reliability in a static manner, citing past reliability estimates is preferable to ignoring it altogether.

The overall calculation of reliability estimates varied in occurrence during the decade. In 1980 the combined estimates (i.e. for established scales, modified/new scales, and inter-rater reliability coefficients) were calculated on 21% (39) of the scales, which was larger than the 19% (35) that cited past reliability estimates. In 1985 there were reports of the combined estimates for 37% (76) of the scales, which again was larger than the 20% of citation of past reliability estimates. In 1990 the upward trend appeared to have waned as noted by 32% (63) of the scales being accompanied by sample reliability reports versus 42% of the scales being accompanied with dated reliability estimates. While there is a general upward trend, the decline from 1985 to 1990 may be an indication that interest in obtaining current estimates of reliability has reached a plateau.

As predicted, there was an increase of the number of sample reliability estimates with previously developed scales. In 1980 only 5% (9) of the authors reported calculating the reliability estimates for the sample, while in 1990 18% (36) had done so. This tendency may be the strongest indication that more researchers are viewing reliability as residing in the data as opposed to being an inherent characteristic of an instrument. There was also an increase in the calculation of reliability estimates for scales which were new and/or modified. Admittedly, researchers who calculated the reliability

estimates for new scales or modified scales could have done so without perceiving the need to obtain reliability measures in for future samples.

In conclusion, the need to report reliability estimates remains great. Authors' tendencies to discuss reliability estimates are growing; however, the reliance on past reliability estimates appears to be strong. While fluid views of reliability appear to be increasing, static views seem to be the prevalent perspective. It is encouraging to note that an upward trend in the calculation of sample reliability estimates does exist. The increase in reporting estimates of reliability for each sample will hopefully increase the clarity of findings by limiting the number of confounding factors.

## References

Allen, M. J. & Yen, W. M. (1979). Introduction to measurement theory. Belmont, CA: Brooks/Cole.

Dawis, R. V. (1987). Scale construction. Journal of Counseling Psychology, 34, 481-489.

Eason, S. (1991). Why generalizability theory yields better results than classical test theory: A primer with concrete examples. In B. Thompson (Ed.), Advances in Educational Research: Substantive Findings, Methodological Developments (Vol. 1, pp. 83-98). Greenwich, CT: JAI Press.

Hains, A. A. & Szyjakowski. (1990). A cognitive stress-reduction intervention program for adolescents. Journal of Counseling Psychology, 37, 79-84.

Kazdin, A. E. (1980). Research design in clinical psychology. New York: Harper & Row:

Locke, L. F., Spirduso, W. W., & S. J. (1987). Proposals that work: A guide for planning dissertations and grant proposals (2nd ed.). Newbury Park, CA: Sage.

Long, B. C. (1990). Relationship between coping strategies, sex-typed traits, and environmental characteristics: A comparison of male and female managers. Journal of Counseling Psychology, 37, 185-194.

Meier, S. T. & Davis, S. R. (1990). Trends in reporting psychometric properties of scales used in counseling psychology research. Journal of Counseling Psychology, 37, 113-115.

Melancon, J. G. & Thompson, B. (1990). Measurement characteristics of the MAA math placement tests. (ERIC Document Reproduction Service No. ED 325 510)

Nunnally, J. C. (1972). Reliability of measurement. In H. E. Mitzel (Ed.), Encyclopedia of educational research (pp. 1589-1601). New York: Free Press.

Robbins, & Haase. (1985). Power of nonverbal cues in counseling interactions: Availability, vividness, or salience? Journal of Counseling Psychology, 32, 502-513.

Shavelson, R. J. & Webb, N. M. (1991). Generalizability theory: A primer. New York: Sage.

Slaney, R. B. (1980). Expressed vocational choice and vocational indecision. Journal of Counseling Psychology, 27, 122-129.

Snyder, P., Lawson, S., Thompson, B., Stricklin, S., & Sexton, D. (in press). Evaluating the psychometric integrity of instruments used in early intervention research: The Battelle Developmental Inventory. Topics in Early Childhood Special Education.

Thompson, Bruce. (1992). Two and one-half decades of leadership in measurement and evaluation. Journal of Counseling & Development, 70, 434-438.

Willson, V. L. (1980). Research techniques in AERJ articles: 1969 to 1978. Educational Researcher, 9(6), 5-10.